A

Synopsis on

An Artificial Intelligence Based Approach to Efficient Data Retrieval in Big Data Analytics

For Partial Fulfillment of Registration

In

DOCTOR OF PHILOSOPHY

In

Computer Science and Applications

Submitted By:
Divya
Research Scholar
Dept. of Computer Science & Applications

Co-Supervisor
Naveen Dahiya
Associate Professor
Computer Science

Supervisor:
Vikram Singh
Professor
Computer Science

CONTENTS

1. **INTRODUCTION**
   - **1.1 Big data**
   - **1.2 Big Data Analytics**
   - **1.3 Artificial Intelligence**

2. **RELATED WORK**
   - **2.1 Research Gaps: Big Data Analytics Challenges and Barriers**

3. **SCOPE AND RATIONALE OF STUDY**

4. **PROBLEM STATEMENT, OBJECTIVES AND RESEARCH METHODOLOGY**
   - **4.1 Problem Statement**
   - **4.2 Objectives of the study & Research methodology**

5. **TENTATIVE CHAPTER SCHEME**

6. **REFERENCES**

1. Introduction

The Internet has become an inescapable source of information for users, enabling access to a large and increasing amount of information and characterized by an exponential growth of Internet users with the number of Internet users being maximum in Asia, followed by Europe and North America. Nowadays, in the 4th Industrial revolution era, organizations and governments focus on the development of capabilities that provide knowledge extracted from large and complex data sets, commonly known as “Big Data”.

Big Data is essentially extremely large and/or complex data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially related to human behavior and interactions. Big data is a buzzword in the last years in the business and economics fields, since it plays an essential role in economic activity and has strengthened its role in creating economic value by enabling new ways to spur innovation and productivity growth.

With the advent of the Web 2.0 and Web 3.0 users are interacting with each other on social media platforms like YouTube, Facebook or Twitter on a scale much greater than ever before. With increased connectivity, multi-modal access, and the rise of Social Media we are witnessing an exponential growth of User Generated Content. The number of Internet users has risen at an accelerated pace, accompanied by the evolution of devices. Today, these provide near full-time access to the Internet. The main contributor to this new reality is the rise of mobile technology, both concerning the devices themselves and mobile data connectivity. This allows an increasing number of Internet users to access the Internet from virtually anywhere and at any time. This increased connectivity has also enabled companies to access and collect big amounts of data in an easier, cheaper and faster manner. Mobile devices like smart phones and tablets being both Internet enabled and mobile have contributed to this paradigm to a greater extent.

Digital enterprises like Google, Amazon and Facebook highlight the significance of big data, indicating the various ways that can be used from predictive analytics, user behaviour analytics , or certain other advanced data analytics methods that extract valuable information from data. Analysis of data sets could find new correlations to spot business trends giving a competitive edge over market players, prevent diseases, combat crime, etc. Nonetheless, exploitation of big data needs people with skills and expertise who will able to capture value from data insights providing significant knowledge to managers and decision-makers.

As information progressively shifts from a physical medium to online content, making sense of this information is crucial to provide the best set of information resources to users. The increasing ability to access information was accompanied by the added speed and reach in information sharing. This development poses a significant impact in social,

economic and political contexts, to which the advent of social media platforms played and continues to play a key role. The creation and proliferation of such platforms allow users to post information in real-time and interact with others. This provoked an increase of available information known as social media data, and a growing demand concerning the computational capability to analyze, interpret and act upon such information. Given the scale, the dynamics, the heterogeneity of the content and the decentralization of the contents provided by the individual users owing to the use of different technology and platforms it has become very important to understand and identify the contents that will gain popularity for efficient maintenance and growth of such User Generated Content sites.

The User Generated Content may be categorized as textual data or visual data (images and videos). All these contents play a very important role in formation of opinion in the modern information technology age. Digital images and video are becoming an integral part of human communications. The ease of creating and capturing digital imagery has enabled its proliferation, making our interaction with today's on-line information sources largely a "visual" one. We are increasingly using visual content in expressing ideas, reporting news, educating and entertaining. But, with the growing amount of digital visual media, how do we effectively search for visual information? Can solutions be developed that have the same effectiveness as the existing text and non-visual information search engines? With the increasing numbers of distributed repositories and users, how do we design scalable visual information retrieval systems?

1.1 Big Data

The term "Big data" has been in use since 1990s and it refers to large datasets that are not able to be captured, stored, managed, processed and analysed by typical software tools (Manyika, J., et al.:,2011). These data sets that are huge -not only in size- but also in heterogeneity and complexity, containing structured, semi-structured and unstructured data. In addition, big data includes data that comes in several formats including text, sound, video, image and more. This unstructured data is growing faster than structured and have captured the 90% of all the data (Gantz, J,2011). The number of Internet users has risen at an accelerated pace, accompanied by the evolution of devices. Today, these provide near full-time access to the Internet. The main contributor to this new reality is the rise of mobile technology, both concerning the devices themselves and mobile data connectivity. This allows an increasing number of Internet users to access the Internet from virtually anywhere and at anytime.

The rapid advancements in wireless communication technology has in the recent years brought an enormous increase in the popularity of User Generated Content (UGC) sites such as YouTube, Facebook, Twitter or Instagram. With the growing connectivity between users, these online social networks have become the most important platform for both information dissemination and information search by users, individuals and

organizations. Therefore, new forms of processing capabilities are required for getting data insights that lead to better decision making.

Basically, Big data can be characterized by the seven Vs: volume, variety, veracity, velocity, variability, visualization and value.

Volume: It refers to the incredibly large size of the data that is generated each second. There are many enabling technologies that have contributed to the proliferation of Big Data such as Internet of Things (IoT), Information and Communication Technologies (ICTs), Artificial Intelligence (AI) etc. Therefore, the data is now even stored in a distributed manner with the organisations having their servers located at the different parts of the globe. Collecting and analysing this data is clearly a big engineering challenge in itself.

Variety represents the increasing diversity of data generation sources and data formats. We no longer have structured data that can fit easily and nicely in a tabular form. Today almost 80% of the total data generated and available is unstructured and with the growth of web and social media networks we have seen an enormous increase in the popularity of User Generated Content (UGC) sites such as YouTube, Facebook, Twitter or Instagram. These platforms are a source of generation and dissemination of billions of hours of video data having not only variety in terms of the content but format, technology used for creation, underlying hardware technology and so on. From messages, updates, photos and videos that are posted in these social media networks, SMS, GPS signals from smartphones, customers transactions in banking, e-business and retail, voice data in call centres etc. Many of the crucial sources of big data are comparatively novel. So we can say that big data differs in data types that are generated, consisting of structured data (tables, records), unstructured data (text and voice), semi-structured data (XML, RSS feeds) and other data that is difficult to classify like data deriving from audio, video and other appliances.

Variability is often confused with variety, but variability is related with rapid change of meaning. For instance, words in a text can have a different meaning according to context of a text, thus for an accurate sentiment analysis, algorithms need to find out the meaning (sentiment) of a word taking into account the whole context.

Velocity. Big data is characterized by the high speed of data generation. The rapid advancements in wireless communication technology has in the recent years brought an enormous increase in data generated by connected devices and with the growing connectivity between users, the web has become the most important platform for data generation, information dissemination and information search and/or retrieval by users, individuals and organizations.This speed is extremely significant for enterprises in taking various actions that enable them to be more agile, gaining competitive advantage against competitors. Despite the fact that some enterprises have already exploited big data (click-streams data) to offer their customers purchase recommendations, nowadays enterprises

though big data analytics have the ability to analyse and understand data taking actions in real-time.

Veracity of data refers to the data reliability and accuracy. Given the scale, the dynamics, the heterogeneity of the content and the decentralization of the contents provided by the individual users owing to the use of different technology and platforms it is very challenging to accurately distinguish the reliable from the unreliable, and accurate from inaccurate. For instance, if we are dealing with twitter posts with hashtags, abbreviations and typos, cleaning the data is a task. Even the data is reliable or not is a questionable thing in such a scenario. The data collection has data that are not clean and accurate, thus data veracity refers to the data uncertainty and the level of reliability correlated with some type of data.

Visualization. Data visualization is the way of visually representing data. Both the qualitative and quantitative information can be represented schematically indicating various patterns, trends, anomalies etc. In order to make sense of this large amount of data, the visual representation has to be effective so that one can leverage and put this large amount of data to the benefit of enterprises, organizations, communities and consumers.

Value: value refers to the worth of the data being extracted. Only when the data could be put in some use, it is beneficial else it is totally useless. A cost benefit analysis is important to be understood to ensure the fruitful usage of data.

1.2 Big Data Analytics

Data science is defined as a collection of fundamental principles that promotes taking information and knowledge from data (Provost et. al., 2013). Over the years, it has been found that for any enterprise the data driven approaches like Business Intelligence (BI) and Business Analytics(BA) are practically indispensable.

"Big Data Analytics" refers to advanced analytic techniques associated with the analysis of large data-sets in enterprises and includes data science, business intelligence and business analytics. Using advanced technologies, Big Data Analytics (BDA) includes data management, open-source programming like Hadoop, statistical analysis like sentiment and time-series analysis, visualization tools that help structure and connect data to uncover hidden patterns, undiscovered correlations and other actionable insights. The process of BDA is important for strategic decision making leading to competitive edge over rivals. In that context, the process of getting insights from big data can be divided into two phases: data management and data analysis. Data management is related with the processes and technologies for data generation, storage, mining and preparation for analysis, while data analysis refers to the methods and techniques for analysis and interpretation of the insights coming from big data (Burstein, F. et. al.,2008)

Business Intelligence(BI) is defined as a the methodologies, systems and applications for collecting, preparing and analysing data to provide information helping decision makers. In other words, BI systems are data-driven decision making systems (Larson, D. et. al.,2016), while Business Analytics(BA) basically streamlines vast amount of data to enhance its value. Thus it encompasses techniques, technologies, systems and applications that are used to analyse critical business data for supporting them to understand their business environment and take business decisions on time.

The ongoing increase of mobile devices against the number of computers introduced a new era of business analytics, including the analysis of user-generated content by social media channels. Mobile devices have the capability to promote e.g. highly mobile, location-aware and person-centred processes and transactions. Therefore, Data-driven decision making is on data coming from all the sources of enterprises, while predictions and machine learning are based on traditional data and new innovative sources like IoT and AI.

Analytics can be divided into four categories, namely descriptive analytics, diagnostic analytics, predictive analytics and prescriptive analytics.

Descriptive analytics is based on historical and current data and provides insights about what happened in the past and the correlations between various determinants identifying patterns using statistical measures like mean, range and standard deviation. For instance, techniques like online analytical processing (OLAP) exploits knowledge from the past experience to provide answers in what's happening in the organizations. Common examples of descriptive analytics include data visualization, dashboards, reports, charts and graphs presenting key metrics of enterprises including sales, orders, customers, financial performance etc. To get the diagnostic analytics, the analytics tool employs techniques like drill-down, deep learning, data discovery, correlations, etc.

Diagnostic analytics based also in historical data provide insights about the root-cause of some outcomes of the past. Thus, organizations can take better decisions avoiding errors and negative results of the past.

Predictive analytics is about forecasting and providing an estimation for the probability about what might occur, for example, future locations of the subscribers, future traffic pattern and network congestion, etc. for defining opportunities or risks in the future. It uses various statistical techniques such as machine learning, data mining, data modelling and machine learning as some analytical process and game- theoretic analysis. One of the most known applications of that type of analytics is the prediction of customer behaviour, determining operations, marketing and preventing risk. Using historical and other available data, predictive analytics are able to uncover patterns and identify relationships in data that can be used for forecasting (Davenport, T. , 2014). Predictive analytics is thus an important tool for organizations to have a competitive edge over the

rest by identify future trends and patterns, presenting innovative products/services and innovations in their business models.

Prescriptive analytics provide one step further to not only predict the future events but also to suggest decision options for slicing (i.e., how to slice, how many slices), virtualization, edge-computing, etc., along with the implications of each decision option. Therefore, the prescriptive analytics need a useful predictive model, actionable data and a feedback system for tracking down the results generated by the action taken.

Prescriptive analytics using high level modelling tools is able to contribute remarkably to the performance and efficiency of organizations, through smarter and faster decision with lower cost and risk and identifying optimal solutions for resource allocation (Gandomi, A., et. al, 2015). Thus we can easily grasp the role that an effective predictive and prescriptive model can play in strategic decision making of organisations.

1.3 Artificial Intelligence

Artificial intelligence may be referred to as the ability of a machine to learn from experience, adjust to new inputs and perform human-like tasks. With the advancements of Big Data technologies, AI is gaining popularity to solve many of the challenges faced by the community for the manipulation of the Big Data and has provided for enhanced decision making, reinventing business models and ecosystems, popularity prediction etc. the various approaches/ domains of AI that serve as a probable candidate for Big Data Analytics are as follows:

To manage Big Data for providing predictive analysis and providing suggestions, Machine Learning and Artificial Intelligence, both are emerging as a good solution. Machine Learning is excellent for predictive analytics , Artificial Intelligence goes beyond predictions and prescribe plans/suggestions with implications to realize a benefit. In fact ML is developed mainly from AI, hence the two overlap. ML has tools to extract relevant information, suggestions, and predictions from the data sets that are too large and too complex. AI has a broader scope: to replicate human intelligence or some aspects of it and other cognitive functions. There are some other approaches also that may be used to handle, manage and manipulate Big Data. Some of them are discussed below.

Genetic algorithm: It is a metaheuristic approach inspired by the process of natural selection that belongs to the larger class of evolutionary algorithms (EA) introduced by John Holland in 1960 based on the concept of Darwin's theory of evolution.. Genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover and selection. Genetic Algorithms can be used effectively in feature selection that are important in prediction of the target variable.

Neural Networks: A neural network is a powerful computational data model that is able to capture and represent complex input/output relationships and perform tasks mimicking the behaviour of a human brain. A neural network acquires knowledge through learning. The true power and advantage of neural networks lies in their ability to represent both linear and non-linear relationships and in their ability to learn these relationships directly from the data being modelled. Traditional linear models are simply inadequate when it comes to modelling data that contains non-linear characteristics. Thus neural networks can be used in various applications like Process Modelling and Control ,Machine Diagnostics ,Portfolio Management - Target Recognition -.Medical Diagnosis - Credit Rating - Targeted Marketing - Voice Recognition, Financial Forecasting - Quality Control, Intelligent Searching ,Fraud Detection

Fuzzy based learning for big data mining: The term *fuzzy logic* was introduced with the 1965 proposal of fuzzy set theory by Lotfi Zadeh (Zadeh, 1965). Fuzzy logic is a form of many-valued logic in which the truth values of variables may be any real number between 0 and 1 inclusive. It is employed to handle the concept of partial truth, where the truth value may range between completely true and completely false. By contrast, in Boolean logic, the truth values of variables may only be the integer values 0 or 1.It is based on the observation that people make decisions based on imprecise and non-numerical information, fuzzy models or sets are mathematical means of representing vagueness and imprecise information, hence the term fuzzy. These models have the capability of recognising, representing, manipulating, interpreting, and utilising data and information that are vague and lack certainty. Fuzzy logic has been applied to many fields, from control theory to artificial intelligence in real life scenarios where the data sets are large, heterogenous and uncertain.

2. Related Work

(Chen and Zhang, 2014) demonstrated a detailed view of Big Data, including Big Data applications, Big Data opportunities and challenges, as well as the state-of-the-art techniques and technologies currently adopted to deal with the Big Data challenges. Various underlying methodologies to handle the data deluge, for example, granular computing, cloud computing, bio-inspired computing, and quantum computing are also discussed in detail.

The authors in (Gandomi and Haider, 2015) presented a description of big data from the view point of practitioners and academics, with the main focus being on the analytic methods used for big data, primarily dealing with unstructured data. Authors highlighted the need to develop appropriate and efficient analytical methods to leverage massive volumes of heterogeneous data in unstructured text, audio, and video formats. The Authors reiterated the need to devise new tools for predictive analytics for structured big data. The need of the hour is to develop computationally efficient algorithms that may avoid big data pitfalls which are contributed to heterogeneity, noise, and the massive size.

In their research (Hashem et. al., 2015) presented the definition, characteristics, and classification of big data along with an overview of cloud computing. The relationship between cloud computing and big data is identified and the authors discussed the storage systems of big data and background of Hadoop technology. Research challenges with main focus on scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy, legal and regulatory issues, and governance are discussed. Lastly, open research issues that require substantial research efforts are summarized.

(Yan et. al., 2016) introduced a Convolution Neural Network based semantic indexing method for biomedical abstract document collections. The data is categorized into coarse subsets by clustering. Next, a high-dimensional space representation with Wikipedia category extension, which contains more semantic information than bag-of-words, is constructed. Thereafter, a hierarchical CNN indexing architecture for learning documents from a coarse-to fine-grained level with several multi-label training techniques is designed. Finally, authors performed comparative experiments for semantic indexing of biomedical abstract documents.

The authors in (kaisper et. al., 2013) identified critical issues associated with data storage, management, and processing and initiated a collaborative research effort to begin examining big data issues and challenges. A number of concerns that need to be addressed in the coming decade are also presented and discussed.

(Provost et. al., 2013) and (Brynjolfsson et. al.,2011) identified Data Science and Big Data Technologies as data-driven decision mechanisms aimed at improving decision making process and contribute in better enterprise performance in terms of productivity and profitability.

The authors in (Trzcinski et. al., 2017) have identified that the external context of the content plays an important role in the distribution pattern of video i.e. if the topic of the video is trending on other social media its popularity is also expected to be high. They have introduced Popularity support vector regression method that uses Gaussian Radial Basis Function and provided a comparative analysis of univariate Linear Regression, Multivariate Linear Regression and Multivariate Radial Basis Function The visual and textual cues are taken into consideration.

(Zhang et. al., 2018) have identified that the popular and unpopular videos may have similar early view patterns and due to this reason classifying videos on the basis of early view patterns limits performance benefits. They have thus proposed a model in which they have identified an Early Rating Pattern based on Early views, Early viewers attitude and factored in Multisource data. On this basis they have trained a multilayer perceptron network to identify long term popularity.

In their research (Vassakis, et. al.,2017) identified the impact of Big Data on business and how it has helped businesses evolve. They have also discussed various techniques to analyse critical data gathered through Internet of Things. Now the industry is regarded to be data driven where every sector (agriculture, health, energy and infrastructure, economics and insurance, sports, food and transportation) and every world economy is driven by data. various obstacles to adopt data-driven approach and get valuable knowledge through big data are also discussed in the paper.

(Najafabadi et. al., 2015) discussed how Deep Learning can help in solving specific problems in Big Data Analytics and explored the application of Deep Learning for specific Big Data Analytics, including learning from massive volumes of data, semantic indexing, discriminative tasks, and data tagging. A special attention is paid towards improving Deep Learning methods to sync with challenges associated with Big Data Analytics which includes learning from streaming data, dealing with high dimensionality of data, scalability of models, and distributed and parallel computing. Finally, authors presented some future areas needing innovation in Deep Learning for Big Data Analytics, including data sampling for generating useful high-level abstractions, domain (data distribution) adaption, defining criteria for extracting good data representations for discriminative and indexing tasks, semi-supervised learning, and active learning

The authors in (Yanqing et. al., 2019) identified the challenges associated with the use and impact of revitalised AI based systems for decision making and offered a set of research propositions for information systems (IS) researchers. They have discussed AI for decision making in general and the specific issues regarding the interaction and integration of AI to support or replace human decision makers in particular. To advance research on the use of AI for decision making in the era of Big Data, the paper offers twelve research propositions for IS researchers in terms of conceptual and theoretical development, AI technology-human interaction, and AI implementation.

(Kibriya et. al., 2019) have identified that due to diversified service requirements, heterogenous applications, devices and networks, the next generation wireless systems are evolving as a complex structure. The traditional networking approaches viz., reactive, centrally-managed approach and conventional data analysis tools that have limited capability (space and time) are not competent anymore and cannot satisfy and serve that future complex networks regarding operation and optimization cost-effectively. They have proposed a proactive, self-aware, self- adaptive and predictive networking which makes the system smart, intelligent, and facilitates efficient as well as cost-effective operation and optimization. The paper also discusses the challenges and benefits of adopting big data analytics, machine learning, and artificial intelligence in the next-generation communication systems.

2.1 Research Gaps: Big Data Analytics Challenges and Barriers

Big Data Analytics faces a number of challenges. Some key areas where improvements can be made include: data quality and validation, data cleansing, feature engineering, high-dimensionality and data reduction, data representations and distributed data sources, data sampling, scalability of algorithms, data visualization, parallel and distributed data processing, real-time analysis and decision making, crowd sourcing and semantic input for improved data analysis, tracing and analysing data provenance, data discovery and integration, parallel and distributed computing, exploratory data analysis and interpretation, integrating heterogeneous data, and developing new models for massive data computation. Some of the challenges faced by the industry are enumerated as follows:

As data sets are becoming bigger and more diverse, there is a big challenge to incorporate them into an analytical platform. If this is overlooked, it will create gaps and lead to wrong messages and insights. So there is an urgent need for synchronisation across disparate data sources.

There is a wide gap in the access to huge amount of data and how to gain insight from the data which needs to be bridged in an effective manner. With new technologies being developed every day which technology is suited to which problem is an important parameter.

With the ever increasing nature of data, the storage of this massive amount of data is also becoming a real challenge for everyone. Unlike the structured data, big data comprises of semi structured and unstructured data and thus traditional options like data lakes/ warehouses are not sufficient enough for the task. Missing data, inconsistent data, logic conflicts, and duplicates data all result in data quality challenges.

The Big Data tools used for analysis and storage utilizes the data disparate sources. This eventually leads to a high risk of exposure of the data, making it vulnerable. Thus, the rise of voluminous amount of data increases privacy and security concerns. Enterprises need to use safeguards to ensure that data that is used does not violate a customers' personal privacy. Thus, data policies including privacy, security, intellectual property and liability issues, should be addressed in order to address these concerns.

The traditional systems are not capable enough of performing the analytics on the data, which is constantly in motion. So, there is a need for a cohesive set of new analytical tools and techniques to provide solutions for these areas. In machine learning community, Deep Learning can be used as a new tool that helps in Big Data analytics. Big Data Analytics and Deep Learning are two high-focus of data science.

3. Scope and Rationale of Study

Today we are living in an information age where real- time analytics is a big trend that enterprises need to pay attention at in the near future. Despite the challenges and issues that are addressed, it is proven that analytics-driven management has significant implications on enterprises, whether they are looking for growth, efficiency or competitive differentiation. Therefore, Big data analytics have seemingly unlimited potential to help an enterprise to grow and reveal its data potential.

Some of the interesting application areas of Big Data Analytics include Fraud and Risk Detection, Healthcare sector especially Medical Image Analysis, Genetics & Genomics, Drug Development and discovery, Virtual assistance for patients and customer support, Internet Search, Targeted Advertising, Website Recommendations, Advanced Image Recognition, Speech Recognition, Airline Route Planning to Predict flight delay, Decide which class of airplanes to buy, Whether to directly land at the destination or take a halt in between, Gaming and Augmented Reality and so on. Moreover the potential scenarios for big data analytics fall in the category of Clustering, Classification, Recommendation Systems, Associations, Real time analytics and Social media mining.

Data Analytics can make it easy for the users to filter huge amount of information easily and efficiently. It can be very helpful for content optimization. For instance, news websites select and organize articles from a highly dynamic content pool. Instead of relying on human editors, websites can refine their decisions through automatic solutions using online content optimization methods.

Accurate popularity predictions can boost a site's traffic and publishers can increase the traffic through social media optimization solutions. Faced with an ever increasing traffic demand, content providers and content delivery networks set large scale caching infrastructures to distribute copies of the content across multiple locations. Optimal placement of replicas in terms of location and number of copies under bandwidth and storage constraints depend on how accurate one can predict the future users' demand: which content will be popular, its geographic locality of interest, and the amount of attention it will generate.

Another domain where popularity prediction can prove to be useful is Mobile Data Offloading. Under the increasing consumption of mobile data traffic, telecom operators look for new solutions to reduce the traffic from cellular networks. Opportunistic networks have been proposed as a solution to offload content with non-real time constraints. In this context, a strategy is to take cue from the spatiotemporal mobile user requests, proactively prefetch popular web items into mobile user's cache according to predicted future demand, and rely on device to device communication to deal with future content requests.

4. PROBLEM STATEMENT, OBJECTIVES AND RESEARCH METHODOLOGY

4.1 Problem Statement

Based on the study of the literature, gaps have been identified in the application of Artificial Intelligence in Big Data Analytics. Our research approach will be focused on thorough understanding of underlying concepts, improving in the existing methods and dealing with application of Artificial Intelligence in Big Data Analytics. The problem can be stated as **"An Artificial Intelligence based approach for Efficient Data Retrieval in Big Data Analytics"**

The current research study aims to design and apply a novel modification of existing techniques which will be better in terms of salient features including Time and Space complexity.

4.2 Objectives of the study & Research methodology

The work in the field of big data mining and analytics is exhaustive, vast and is scattered across various journals. Due to the multifaceted nature of big data owing to its various formats- structured, semi structured, unstructured and various types- Text, audio, video and biometrics there is an urgent need for proper analytics so that strategic advantages of the data can be derived. The various factors that made this field to be studied extensively include its vast application areas in almost every domain. The methodology followed for the research is tentatively as follows:

1. To study artificial intelligence and its various allied approaches viz. machine learning, genetic algorithms, fuzzy systems, neural networks – deep neural nets and convolution neural nets for their applications in data mining.
2. To identify the challenges associated with the use and impact of artificial intelligence techniques in big data analytics.
3. To identify various application areas that can be used in big data analytics and to identify the Descriptive, diagnostic, predictive and prescriptive big data intelligence areas.
4. To identify the areas in which data is to be considered: text, image, video or biometric data
5. To improve upon the MBRF function that is used for analytics using mathematical Gaussian Radial Basis function.
6. To evaluate the performance of the proposed method and compare it with the existing technique.

5. TENTATIVE CHAPTER SCHEME

Chapter 1: Introduction

Chapter 2: Review of literature

Chapter 3: Research Methodology

Chapter 4: Issues and Loopholes in Big Data Analytics

Chapter 5: Proposed Artificial Intelligence based Model for information Retrieval

Chapter 6: Implementation of proposed AI based Model for efficient information Retrieval

Chapter 7: Performance comparison of Proposed Model with existing Techniques

Chapter 8: Conclusion and future scope

REFERENCES